SignalP 5.0: improved signal peptide predictions across the tree of life using deep neural networks

Jose Juan Almagro Armenteros, Konstantinos D. Tsirigos, Casper Kaae Sønderby, Thomas Nordahl Petersen, Ole Winther, Søren Brunak, Gunnar von Heijne and Henrik Nielsen

Department of Bio and Health Informatics, Technical University of Denmark, Kgs Lyngby, Denmark

Introduction

Signal peptides (SPs) are found in a large number of nascent polypeptide chains, signaling protein export in Bacteria, Archaea and Eukaryotes. The SP possesses a modular architecture with three distinct regions: an N-terminal positively charged region (n-region); a core, hydrophobic region (h-region); and a c-region that contains mostly small and uncharged residues and ends at a characteristic proteolytic cleavage site. During or after translocation through the membrane, an enzyme removes the SP. The enzyme is called Signal Peptidase I or LepB in Bacteria, and orthologs of it are found in Archaea as well as in Eukaryotes. This pathway is termed the general secretion pathway (Sec). In Bacteria and Archaea, there is an alternative export pathway named Tat, as well as an alternative signal peptidase named Signal Peptidase II or Lipoprotein Signal Peptidase (LspA). SignalP was one of the first methods for SP prediction to become available online, being launched as a web-server in 1996.

Mayor improvements since SignalP4

- We present a major update to the SignalP prediction method. SignalP 5.0 [1] is a deep Neural Network-based method combined with a Conditional Random Field.
- We trained and tested the networks on four types of organisms (Eukaryotes, Archaea, Gram-positive bacteria, and Gram-negative bacteria).
- We predict four types of proteins: Sec, Lipo, Tat and 'Other', i.e. globular proteins without SP and transmembrane (TM) proteins with a TM segment within the first 70 amino acids.
- We benchmarked SignalP 5.0 against SignalP 4.1 [2] and 18 other signal peptide prediction algorithms - as many as currently are available either as web-servers or standalone packages.

Dataset

Table 1: The composition of the datasets that was used during the development of SignalP 5.0.

Type	Archaea	Eukaryotes	Gram-negative bacteria	Gram-positive bacteria
Sec signals	66	2,614	509	189
Lipo signals	28	–	1,063	449
Tat signals	–	–	334	95
Other	122	14,659	422	190
Total	216	17,273	2,328	923

SignalP 5.0 model

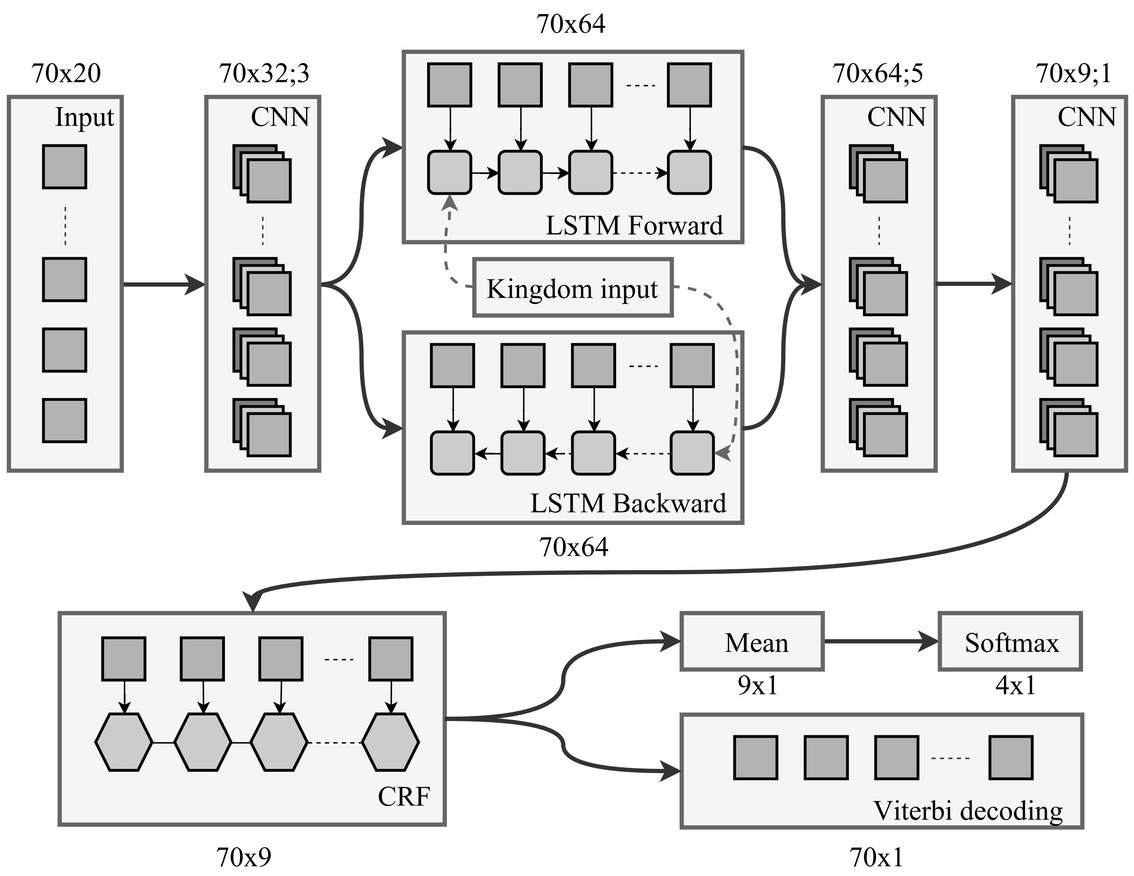

Figure 1: Model architecture representation.

The deep learning model is composed of three primary components:

- One-dimensional convolutions, capturing short range correlations.
- Bidirectional Long-Short Term Memory (LSTM) cells capturing long range sequence dependencies
- A Conditional Random Field (CRF) for predicting the class labels.

To improve performance in kingdoms with little data (notably Archaea), SignalP 5.0 utilizes **transfer learning** between kingdoms. Also, it employs **multimodal learning**, predicting both the individual amino acid labels as well as the global signal peptide type.

Model performance

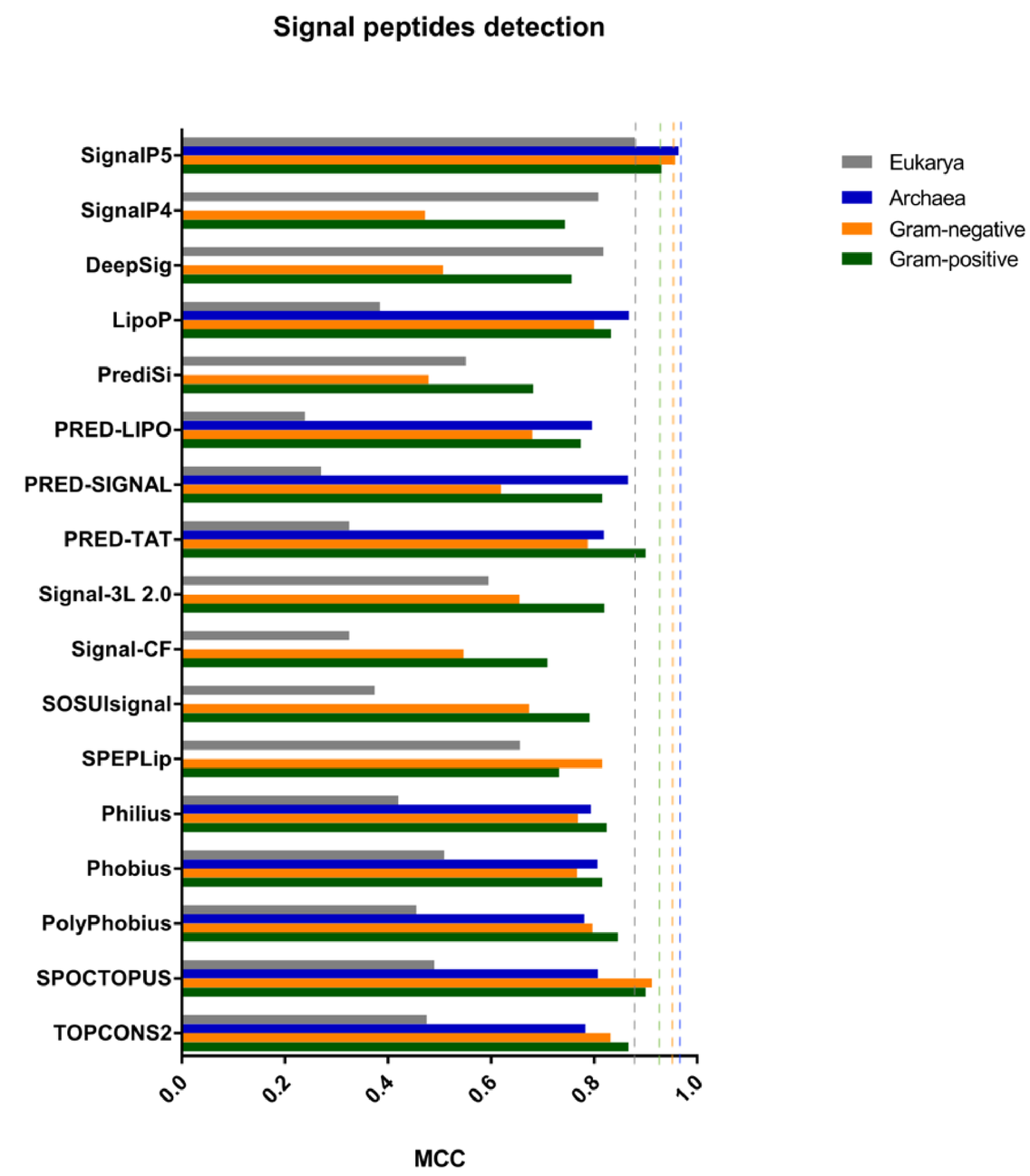

Figure 2: Comparison of signal peptide prediction algorithms performance shown as the Matthews Correlation Coefficient for Eukaryotes, Archaea, Gram-negative bacteria and Gram-positive bacteria.

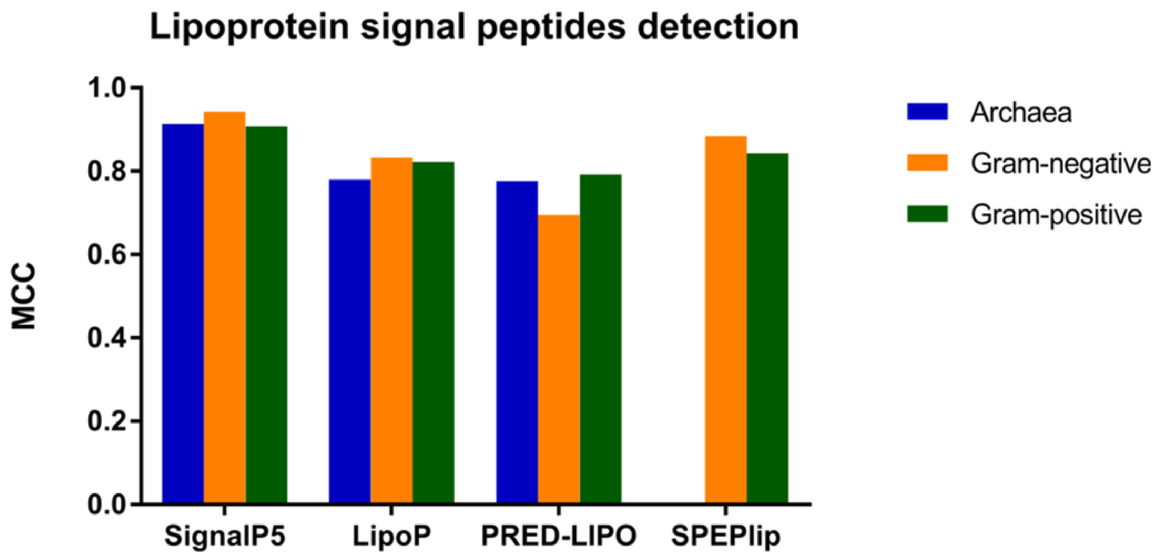

Figure 3: Comparison of prediction algorithms performance on Lipoprotein (top) and Tat (bottom) signal peptides.

References

[1] J. J. Almagro Armenteros, K. D. Tsirigos, C. K. Sønderby, T. N. Petersen, O. Winther, S. Brunak, G. von Heijne, and H. Nielsen. Signalp 5.0: improved signal peptide predictions across the tree of life using deep neural networks. *Under review*, 2018.

[2] T. N. Petersen, S. Brunak, G. von Heijne, and H. Nielsen. Signalp 4.0: discriminating signal peptides from transmembrane regions. *Nature methods*, 8(10):785, 2011.

jjaa@bioinformatics.dtu.dk